A Survey of Artificial Intelligence for Unsupervised Learning

[1]Vrushang Prakashkumar Anand, [2]Vaibhav V. Patel
[1]Student, [2]Lecturer
[1]St Xavier's High School, Surat, India.

Abstract: A major goal of unsupervised learning is to get data representations that are useful for subsequent tasks, without access to supervised labels during training. Typically, this involves minimizing a surrogate objective, like the negative log likelihood of a generative model, with the hope that representations useful for subsequent tasks will arise as a side effect. during this work, we propose instead to directly target later desired tasks by meta-learning an unsupervised learning rule which results in representations useful for those tasks. Specifically, we target semi-supervised classification performance, and that an algorithm – an unsupervised weight update rule – that produces representations useful for this task. Additionally, we constrain our unsupervised update rule to a be a biologically-motivated, neuron-local function, which enables it to generalize to different neural network architectures, datasets, and data modalities. We show that the meta-learned update rule produces useful features and sometimes outperforms existing unsupervised learning techniques. We further show that the meta-learned unsupervised update rule generalizes to coach networks with different widths, depths, and nonlinearities. It also generalizes to coach on data with randomly permuted input dimensions and even generalizes from image datasets to a text task**.**

IndexTerms **– Artificial Intelligence, Supervised Learning, Neural Network, Architectures, Datasets.**

I. INTRODUCTION

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.[1] It infers a function from labeled training data consisting of a set of training examples.[2] In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way (see inductive bias). The parallel task in human and animal psychology is often referred to as concept learning.

Unsupervised learning is a type of self-organized Hebbian learning that helps find previously unknown patterns in data set without pre-existing labels. It is also known as self-organization and allows modeling probability densities of given inputs.[1] It is one of the main three categories of machine learning, along with supervised and reinforcement learning. Semi-supervised learning has also been described, and is a hybridization of supervised and unsupervised techniques.

Two of the main methods used in unsupervised learning are principal component and cluster analysis. Cluster analysis is used in unsupervised learning to group, or segment, datasets with shared attributes in order to extrapolate algorithmic relationships.[2] Cluster analysis is a branch of machine learning that groups the data that has not been labelled, classified or categorized. Instead of responding to feedback, cluster analysis identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. This approach helps detect anomalous data points that do not fit into either group.

A central application of unsupervised learning is in the field of density estimation in statistics,[3] though unsupervised learning encompasses many other domains involving summarizing and explaining data features. It could be contrasted with supervised learning by saying that whereas supervised learning intends to infer a conditional probability distribution {\textstyle p_{X}(x\,|\,y)}{\textstyle p_{X}(x\,|\,y)} conditioned on the label {\textstyle y}{\textstyle y} of input data; unsupervised learning intends to infer an a priori probability distribution {\textstyle p_{X}(x)}{\textstyle p_{X}(x)}.

II. RELATED WORK

2.1 SCALING UP REPRESENTATION LEARNING FOR NATURAL LANGUAGE

Learning representations of natural language has been shown to be useful for a wide range of NLP tasks and has been widely adopted (Mikolov et al., 2013; Le & Mikolov, 2014; Peters et al., 2018; Devlin et al., 2019; Radford et al., 2018; 2019). One of the most significant changes in the last two years is the shift from pre-training word embeddings, whether standard (Mikolov et al., 2013; Pennington et al., 2014) or contextualized (McCann et al., 2017; Peters et al., 2018), to full-network pre-training followed by task-specific fine-tuning (Radford et al., 2018; Devlin et al., 2019). In this line of work, it is often shown that larger model size improves performance. For example, Devlin et al. (2019) show that across three selected natural language understanding tasks, using larger hidden size, more hidden layers, and more attention heads always leads to better performance. However, they stop at a hidden size of 1024. We show that, under the same setting, increasing the hidden size to 2048 leads to model degradation and hence worse performance. Therefore, scaling up representation learning for natural language is not as easy as simply increasing model size. In addition, it is difficult to experiment with large models due to computational constraints, especially in terms of GPU/TPU memory limitations. Given that current state-of-the-art models often have hundreds of millions or even billions of parameters, we can easily hit memory limits. To address this issue, Chen et al. (2016) propose a method called gradient checkpointing to reduce the memory requirement to be sublinear at the cost of an extra forward pass. Gomez et al. (2017) propose a way to reconstruct each layer's activations from the next layer so that they do not need to store the intermediate activations. Both methods reduce the memory consumption at the cost of speed. In contrast, our parameter-reduction techniques reduce memory consumption and increase training speed.

2.2 CROSS-LAYER PARAMETER SHARING

The idea of sharing parameters across layers has been previously explored with the Transformer architecture (Vaswani et al., 2017), but this prior work has focused on training for standard encoderdecoder tasks rather than the pretraining/finetuning setting. Different from our observations, Dehghani et al. (2018) show that networks with cross-layer parameter sharing (Universal Transformer, UT) get better performance on language modeling and subject-verb agreement than the standard transformer. Very recently, Bai et al. (2019) propose a Deep Equilibrium Model (DQE) for transformer networks and show that DQE can reach an equilibrium point for which the input embedding and the output embedding of a certain layer stay the same. Our observations show that our embeddings are oscillating rather than converging. Hao et al. (2019) combine a parameter-sharing transformer with the standard one, which further increases the number of parameters of the standard transformer.

2.3 SENTENCE ORDERING OBJECTIVES

ALBERT uses a pretraining loss based on predicting the ordering of two consecutive segments of text. Several researchers have experimented with pretraining objectives that similarly relate to discourse coherence. Coherence and cohesion in discourse have been widely studied and many phenomena have been identified that connect neighboring text segments (Hobbs, 1979; Halliday & Hasan, 1976; Grosz et al., 1995). Most objectives found effective in practice are quite simple. Skipthought (Kiros et al., 2015) and FastSent (Hill et al., 2016) sentence embeddings are learned by using an encoding of a sentence to predict words in neighboring sentences. Other objectives for sentence embedding learning include predicting future sentences rather than only neighbors (Gan et al., 2017) and predicting explicit discourse markers (Jernite et al., 2017; Nie et al., 2019). Our loss is most similar to the sentence ordering objective of Jernite et al. (2017), where sentence embeddings are learned in order to determine the ordering of two consecutive sentences. Unlike most of the above work, however, our loss is defined on textual segments rather than sentences. BERT (Devlin et al., 2019) uses a loss based on predicting whether the second segment in a pair has been swapped with a segment from another document. We compare to this loss in our experiments and find that sentence ordering is a more challenging pretraining task and more useful for certain downstream tasks. Concurrently to our work, Wang et al. (2019) also try to predict the order of two consecutive segments of text, but they combine it with the original next sentence prediction in a three-way classification task rather than empirically comparing the two.

III. EXPERIMENTAL DESIGN

Considered methods. All the considered methods augment the VAE loss with a regularize: The β-VAE (Higgins et al., 2017a), introduces a hyper parameter in front of the KL regularize of vanilla VAEs to constrain the capacity of the VAE bottleneck. The AnnealedVAE (Burgess et al., 2017) progressively increase the bottleneck capacity so that the encoder can focus on learning one factor of variation at the time (the one that most contribute to a small reconstruction error). The FactorVAE (Kim & Mnih, 2018) and the β-TCVAE (Chen et al., 2018) penalize the total correlation (Watanabe, 1960) with adversarial training (Nguyen et al., 2010; Sugiyama et al., 2012) or with a tractable but biased Monte-Carlo estimator respectively. The DIP-VAE-I and the DIP-VAE-II (Kumar et al., 2017) both penalize the mismatch between the aggregated posterior and a factorized prior. Implementation details and further discussion on the methods can be found in Appendix B and G.

Considered metrics. The BetaVAE metric (Higgins et al., 2017a) measures disentanglement as the accuracy of a linear classifier that predicts the index of a fixed factor of variation. Kim & Mnih (2018) address several issues with this metric in their FactorVAE metric by using a majority vote classifier on a different feature vector which accounts for a corner case in the BetaVAE metric. The Mutual Information Gap (MIG) (Chen et al., 2018) measures for each factor of variation the normalized gap in mutual information between the highest and second highest coordinate in r(x). Instead, the Modularity (Ridgeway & Mozer, 2018) measures if each dimension of r(x) depends on at most a factor of variation using their mutual information. The Disentanglement metric of Eastwood & Williams (2018) (which we call DCI Disentanglement for clarity) computes the entropy of the distribution obtained by normalizing the importance of each dimension of the learned representation for predicting the value of a factor of variation. The SAP score (Kumar et al., 2017) is the average difference of the prediction error of the two most predictive latent dimensions for each factor. Implementation details and further descriptions can be found in Appendix C.

Data sets. We consider four data sets in which x is obtained as a deterministic function of z: dSprites (Higgins et al., 2017a), Cars3D (Reed et al., 2015), SmallNORB (LeCun et al., 2004), Shapes3D (Kim & Mnih, 2018). We also introduce three data sets where the observations x are stochastic given the factor of variations z: Color-dSprites, Noisy-dSprites and Scream-dSprites. In Color-dSprites, the shapes are colored with a random color. In Noisy-dSprites, we consider white-colored shapes on a noisy background. Finally, in Scream-dSprites the background is replaced with a random patch in a random color shade extracted from the famous The Scream painting (Munch, 1893). The dSprites shape is embedded into the image by inverting the color of its pixels. Further details on the preprocessing of the data can be found in Appendix H.

Inductive biases. To fairly evaluate the different approaches, we separate the effect of regularization (in the form of model choice and regularization strength) from the other inductive biases (e.g., the choice of the neural architecture). Each method uses the same convolutional architecture, optimizer, hyperparameters of the optimizer and batch size. All methods use a Gaussian encoder where the mean and the log variance of each latent factor is parametrized by the deep neural network, a Bernoulli decoder and latent dimension fixed to 10. We note that these are all standard choices in prior work (Higgins et al., 2017a; Kim & Mnih, 2018).

We choose six different regularization strengths, i.e., hyperparameter values, for each of the considered methods. The key idea was to take a wide enough set to ensure that there are useful hyperparameters for different settings for each method and not to focus on specific values known to work for specific data sets. However, the values are partially based on the ranges that are prescribed in the literature (including the hyperparameters suggested by the authors). We fix our experimental setup in advance and we run all the considered methods on each data set for 50 different random seeds and evaluate them on the considered metrics. The full details on the experimental setup are provided in the Appendix G. Our experimental setup, the limitations of this study, and the differences with previous implementations are extensively discussed

IV. LITERATURE REVIEW

Table 4.1: META-LEARNING UPDATE RULES FOR UNSUPERVISED REPRESENTATION LEARNING

TITLE	AUTHOR	PUB. & YEAR	CONCLUSION	FUTURE WORK
Meta-Learning Update Rules For Unsupervised Representation Learning	Luke Metz Niru Maheswaranathan Brian Cheung Jascha Sohl-Dickstein	ICLR 2019	Target semi-supervised classification performance, and we metalearn an algorithm – an unsupervised weight update rule – that produces representations useful for this task. Additionally, we constrain our unsupervised update rule to a be a biologically-motivated, neuron-local function, which enables it to generalize to different neural network architectures, datasets, and data modalities.	We show that the meta-learned update rule produces useful features and sometimes outperforms existing unsupervised learning techniques. We further show that the meta-learned unsupervised update rule generalizes to train networks with different widths, depths, and nonlinearities. It also generalizes to train on data with randomly permuted input dimensions and even generalizes from image datasets to a text task

Table 4.2: A SURVEY OF ARTIFICIAL INTELLIGENCE FOR PROGNOSTICS

TITLE	AUTHOR	PUB. & YEAR	CONCLUSION	FUTURE WORK
A Survey of Artificial Intelligence for Prognostics	Mark Schwabacher Kai Goebel	ICLR 2019	They concluded that prognostics is extremely difficult, and noted that although much research had been done in the area, we were not aware of any deployed prognostic systems that take advantage of measured characteristics of the systems being monitored (but there are of course deployed life usage models). In the two years since then, we have been encouraged to see that more researchers have gotten to the point of building prototype systems that make predictions of remaining useful life. Other researches have built prototype systems that estimate the current level of degradation on a numerical scale, without making the final step of predicting the remaining useful life (Brown et al., 2006.	The question of what to do after detecting a failure precursor. The research in AI planning and scheduling could be very relevant to planning maintenance actions or replanning the mission. Some research has been done in automatically planning the recovery actions to take after diagnosing a failure Verification and validation (V&V). The complexity of AI systems makes them very difficult to verify and validate before deployment. AIbased V&V may offer the potential to help solve this problem. Some research has been done in using the AI approach to verifying diagnostics models

Table 4.3: CHALLENGING COMMON ASSUMPTIONS IN THE UNSUPERVISED LEARNING OF DISENTANGLED REPRESENTATIONS

TITLE	AUTHOR	PUB. & YEAR	CONCLUSION	FUTURE WORK
Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations	Francesco Locatello Stefan Bauer Mario Lucic Gunnar Rätsch Sylvain Gelly Bernhard Schölkopf Olivier Bachem	PMLR 2019	First theoretically show that the unsupervised learning of disentangled representations is fundamentally impossible without inductive biases. We then performed a large-scale empirical study with six state-of-the-art disentanglement	Our study also highlights the need for a sound, robust, and reproducible experimental setup on a diverse set of data sets in order to draw valid conclusions. We have observed that it is easy to draw spurious conclusions from experimental results if one

			methods, six disentanglement metrics on seven data sets and conclude the following: (i) A factorizing aggregated posterior does not seem to necessarily imply that the dimensions in the representation uncorrelated. (ii) Random seeds and hyperparameters seem to matter more than the model but tuning seem to require supervision. (iii) We did not observe that increased disentanglement implies a decreased sample complexity of learning downstream tasks	only considers a subset of methods, metrics and data sets. Hence, we argue that it is crucial for future work to perform experiments on a wide variety of data sets to see whether conclusions and insights are generally applicable.

Table 4.4 : THE LOTTERY TICKET HYPOTHESIS: FINDING SPARSE, TRAINABLE NEURAL NETWORKS

TITLE	AUTHOR	PUB. & YEAR	CONCLUSION	FUTURE WORK
THE LOTTERY TICKET HYPOTHESIS: FINDING SPARSE, TRAINABLE NEURAL NETWORKS	Jonathan Frankle Michael Carbin	ICLR 2019	The initialization that gives rise to a winning ticket is arranged in a particular sparse architecture. Since we uncover winning tickets through heavy use of training data, we hypothesize that the structure of our winning tickets encodes an inductive bias customized to the learning task at hand. Cohen & Shashua (2016) show that the inductive bias embedded in the structure of a deep network determines the kinds of data that it can separate more parameter-efficiently than can a shallow network; although Cohen & Shashua (2016) focus on the pooling geometry of convolutional networks, a similar effect may be at play with the structure of winning tickets, allowing them to learn even when heavily pruned.	The winning tickets we find have initializations that allow them to match the performance of the unpruned networks at sizes too small for randomly-initialized networks to do the same. In future work, we intend to study the properties of these initializations that, in concert with the inductive biases of the pruned network architectures, make these networks particularly adept at learning. On deeper networks (Resnet-18 and VGG-19), iterative pruning is unable to find winning tickets unless we train the networks with learning rate warmup. In future work, we plan to explore why warmup is necessary and whether other improvements to our scheme for identifying winning tickets could obviate the need for these hyperparameter modifications

Table 4.4 : XLNet: Generalized Autoregressive Pretraining for Language Understanding

TITLE	AUTHOR	PUB. & YEAR	CONCLUSION	FUTURE WORK
XLNet: Generalized Autoregressive Pretraining for Language Understanding	Zhilin Yang Zihang Dai Yiming Yang Jaime Carbonell Ruslan Salakhutdinov Quoc V. Le	arXiv:1906.08237v1 [cs.CL] 19 Jun 2019	XLNet is a generalized AR pretraining method that uses a permutation language modeling objective to combine the advantages of AR and AE methods. The neural architecture of XLNet is developed to work seamlessly with the AR objective, including integrating Transformer-XL and careful design of the	we envision applications of XLNet to a wider set of tasks such as vision and reinforcement learning.

			two-stream attention mechanism. XLNet achieves state-of-the-art results various tasks with substantial improvement. In the future,	

REFERENCES

[1] Luke Metz, Niru Maheswaranathan, Brian Cheung, Jascha Sohl-Dickstei, Meta-Learning Update Rules For Unsupervised Representation Learning , ICLR 2019

[2] Mark Schwabacher Kai Goebel, A Survey of Artificial Intelligence for Prognostics , ICLR 2019

[3] Zhilin Yang , Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov , XLNet: Generalized Autoregressive Pretraining for Language Understanding , arXiv:1906.08237v1 [cs.CL] 19 Jun 2019

[4] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly Bernhard Schölkopf Olivier Bachem , Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations , PMLR 2019

[5] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training multi-billion parameter language models using model parallelism, 2019.

[6] Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention. arXiv preprint arXiv:1808.04444, 2018.

[7] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In Neural Information Processing Systems (NeurIPS), 2019.

[8] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. arXiv preprint arXiv:1807.03819, 2018.

[9] Stuart J. Russell, Peter Norvig (2010) Artificial Intelligence: A Modern Approach, Third Edition, Prentice Hall ISBN 9780136042594.

[10] Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar (2012) Foundations of Machine Learning, The MIT Press ISBN 9780262018258.